

## 5.4 Sprint 04

### 5.4.1 Introduction :

La **classification naïve bayésienne** est un type de classification bayésienne probabiliste simple basée sur le [théorème de Bayes](#) avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des [classifieur HYPERLINK](#)

["https://fr.wikipedia.org/wiki/Classifieur\\_lin%C3%A9aire"](https://fr.wikipedia.org/wiki/Classifieur_lin%C3%A9aire)HYPERLINK

["https://fr.wikipedia.org/wiki/Classifieur\\_lin%C3%A9aire"](https://fr.wikipedia.org/wiki/Classifieur_lin%C3%A9aire) HYPERLINK

["https://fr.wikipedia.org/wiki/Classifieur\\_lin%C3%A9aire"](https://fr.wikipedia.org/wiki/Classifieur_lin%C3%A9aire) linéaires.

### Sprint 5: Text Classification using Naive Bayes:

#### Text du projet:

Text Classification is an automated process of classification of text into predefined categories. We can classify Emails into spam or non-spam, news articles into different categories like Politics, Stock Market, Sports, etc. (This sprint is described in my email of text classification and NLP).

Difficulty: Requires basic knowledge in Machine learning, statistics and basic development effort.

### 5.4.2 Plant de travail :

- **Problématique**

Le quatrième sprint de notre projet consiste à classifier les articles obtenus selon leur catégorie d'une façon intelligente, ça veut dire en utilisant un algorithme de classification qui utilise l'intelligence artificiel (Natural language processing), dans notre cas on utilise la naïve bayes.

Pour réussir ce sprint nous devons accomplir les cibles suivantes :

- connaître le NLP et savoir comment l'algorithme de naïves bayes fonctionne
- connecter à la base de données local qui contient les articles
- implémenter naïves bayes classificateur sur les données

## ***1 - c'est quoi NLP***

PNL signifie « Programmation Neuro Linguistique » et existe depuis les années 1970, lorsque ses co-fondateurs, Richard Bandler et John Grinder, ont pour la première fois modelé les thérapeutes Milton Erickson, Gregory Bateson, Fritz Perls et Virginia Satir.

L'ANLP définit la PNL comme un ensemble de modèles, de techniques et de stratégies de modélisation de l'excellence, afin de nous aider à mieux comprendre comment nos processus de pensée et notre comportement, y compris la façon dont le langage que nous utilisons, influencent notre façon de penser et les résultats que nous obtenons. La modélisation de l'excellence dans tous les domaines nous permet d'apporter un changement positif en nous-mêmes et chez les autres.

Une traduction littérale de l'expression `` Programmation Neuro Linguistique " est que la PNL nous habilite, nous permet et nous apprend à mieux comprendre la façon dont notre cerveau (neuro) traite les mots que nous utilisons (linguistiques) et comment cela peut avoir un impact sur notre passé, présent et futur. (Programmation). Cela nous donne des stratégies pour observer le comportement humain et apprendre du meilleur (et du pire) de cela!

En termes simples, le changement est possible - tout ce dont vous avez besoin est un désir de changer et une volonté d'apprendre de nouvelles façons d'être... avec vous-même, vos pensées et avec les autres.

La PNL a été définie comme le « manuel de l'utilisateur pour votre esprit » car l'étude de la PNL nous donne un aperçu de la façon dont nos schémas de pensée peuvent affecter tous les aspects de nos vies.

Dans les années 1970, les Co-créateurs de la PNL définissaient initialement la PNL comme suit:

« La PNL est une attitude qui est une curiosité insatiable sur l'être humain avec une méthodologie qui laisse derrière elle une traînée de techniques.

Richard Bandler (Co-créateur de PNL)

« Les stratégies, outils et techniques de la PNL représentent une opportunité sans pareille pour l'exploration du fonctionnement humain, ou plus précisément, ce sous-ensemble rare et précieux du fonctionnement humain connu sous le nom de génie.

## **2- Naïve bayes :**

Naïve Bayes Classifier est un algorithme populaire en Machine Learning. C'est un algorithme du Supervise Learning utilisé pour la classification. Il est particulièrement utile pour les problématiques de classification de texte. Un exemple d'utilisation du Naïve Bayes est celui du filtre anti-spam.

La naïve Bayes classifier se base sur le théorème de Bayes. Ce dernier est un classique de la théorie des probabilités. Ce théorème est fondé sur les probabilités conditionnelles.

Probabilités conditionnelles : Quelle est la probabilité qu'un événement se produise sachant qu'un autre événement s'est déjà produit.

### **Théorème de Bayes :**

Le théorème de Bayes nous dit que la probabilité d'une hypothèse étant donné une certaine évidence est égale à la probabilité de l'hypothèse multipliée par la probabilité de la preuve donnée à l'hypothèse, puis divisée par la probabilité de la preuve.

$$\Pr(H | E) = \Pr(H) * \Pr(E | H) / \Pr(E)$$

Puisque nous classons des documents, l'« hypothèse » est la suivante: le document entre dans la catégorie C. La « preuve » est constituée des mots W apparaissant dans le document.

Étant donné que les tâches de classification impliquent de comparer deux hypothèses (ou plus), nous pouvons utiliser la forme de ratio du théorème de Bayes, qui compare les numérateurs de la formule ci-dessus (pour les aficionados de Bayes: les temps antérieurs la vraisemblance) pour chaque hypothèse:

$$\Pr(C_1 | W) / \Pr(C_2 | W) = \Pr(C_1) * \Pr(W | C_1) / \Pr(C_2) * \Pr(W | C_2)$$

Puisqu'il y a beaucoup de mots dans un document, la formule devient:

$$\Pr (C_1 | W_1, W_2... W_n) / \Pr (C_2 | W_1, W_2... W_n) =$$

$$\Pr (C_1) * (\Pr (W_1 | C_1) * \Pr (W_2 | C_1) * ... \Pr (W_n | C_1)) /$$

$$\Pr (C_2) * (\Pr (W_1 | C_2) * \Pr (W_2 | C_2) * ... \Pr (W_n | C_2))$$

Par exemple, si je souhaite savoir si un document contenant les mots "préchauffer le four" appartient à la catégorie "livres de cuisine" plutôt que "romans", je compare ceci:

$$\Pr (\text{livre de cuisine}) * \Pr (\text{"préchauffer"} | \text{livre de cuisine}) * \Pr (\text{"le"} | \text{livre de cuisine}) * \Pr (\text{"four"} | \text{livre de cuisine})$$

Pour ça:

$$\Pr (\text{roman}) * \Pr (\text{« préchauffer »} | \text{roman}) * \Pr (\text{« le »} | \text{roman}) * \Pr (\text{« four »} | \text{roman})$$

Si la probabilité qu'il s'agisse d'un livre de cuisine compte tenu de la présence des mots dans le document est supérieure à la probabilité qu'il s'agisse d'un roman, Naïve Bayes renvoie « livre de cuisine ». Si c'est l'inverse, Naïve Bayes renvoie « roman ».

### **Avantages et inconvénients de Naïve Bayes:**

- **Avantages**
- C'est relativement simple à comprendre et à construire
- Il est facile à former, même avec un petit jeu de données
- C'est rapide!
- Il n'est pas sensible aux caractéristiques non pertinentes
- **Inconvénients**
- Il implique que chaque fonctionnalité soit indépendante, ce qui n'est pas toujours le cas.

Les classificateurs de Naïve Bayes sont une famille d'algorithmes reposant sur le principe commun selon lequel la valeur d'une fonctionnalité spécifique est

indépendante de la valeur de toute autre fonctionnalité. Ils nous permettent de prédire la probabilité qu'un événement se produise en fonction de conditions que nous connaissons pour les événements en question. Le nom vient du théorème de Bayes, qui peut être écrit mathématiquement comme suit:

$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$  avec A et B sont les événements et P(B) supérieure strict à 0.

$P(A|B)$  est une probabilité conditionnelle. Plus précisément, c'est la probabilité que l'événement A se produise sachant que B l'événement s'est déjà produit.

Idem,  $P(B|A)$  est une probabilité conditionnelle. Plus précisément, c'est la probabilité que l'événement B se produise sachant qu'A l'événement s'est déjà produit.

$P(A)$  et  $P(B)$  sont les probabilités des événements A et B indépendamment les uns des autres.

Si vous en savoir plus sur les algorithmes de classificateur de Naïve Bayes et de toutes les utilisations du théorème de Bayes, un simple cours de probas suffirait.

- **Implémentation des classificateurs naïves bayes sur les données :**
- **Importation des bibliothèques :**

```
import pandas as pd
import sqlalchemy as sql
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
```

Figure 5.4.1 : capture de code

- **Connection a la base de donnée local :**

```
connect_string = 'mysql://karim:karim@192.168.1.226:3306/karim'

sql_engine = sql.create_engine(connect_string)

query = query = "select * from karim_article"
dataset = pd.read_sql_query(query, sql_engine)
|
```

Figure 5.4.2 : capture de code de connectio a DB

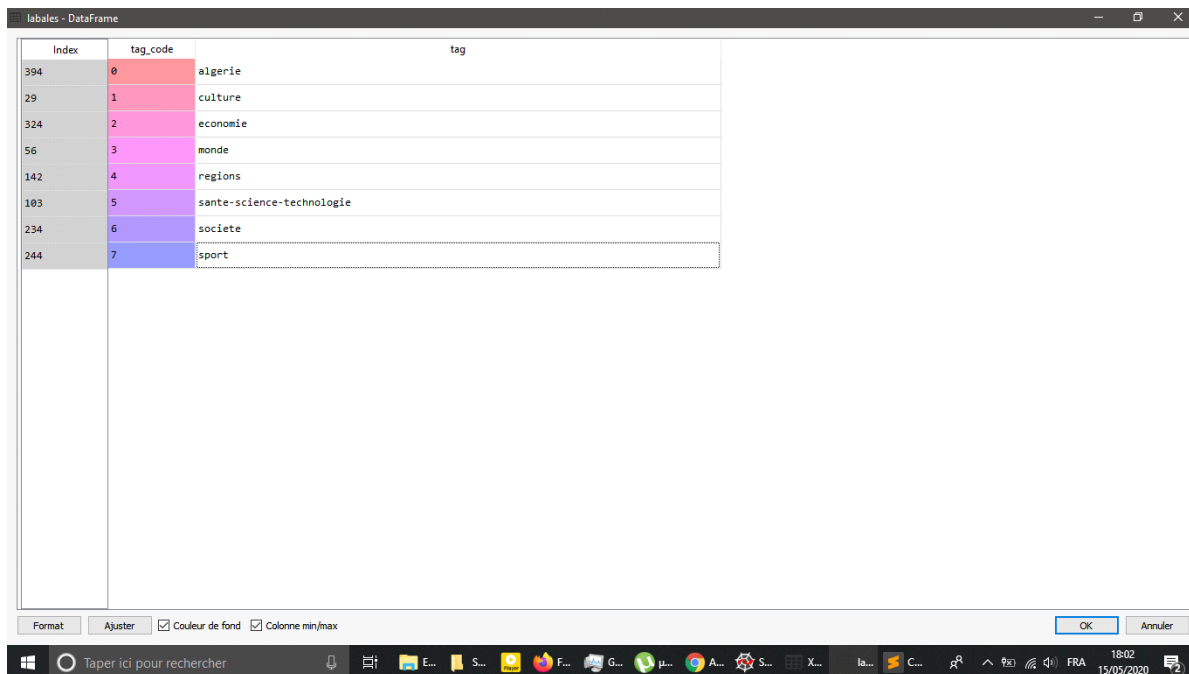
- Etiquetage des données :

```
#1. Prepare the data
lb_make = LabelEncoder()
dataset['tag_code'] = lb_make.fit_transform(dataset['tag'])

labales=dataset[['tag_code','tag']]
labales.sort_values("tag_code", inplace = True)
labales.drop_duplicates(subset = "tag", keep = 'last', inplace = True)
```

Figure 5.4.3: capture de code d'étiquetage

- Résultat :



Index	tag_code	tag
394	0	algerie
29	1	culture
324	2	economie
56	3	monde
142	4	regions
103	5	sante-science-technologie
234	6	societe
244	7	sport

Figure 5.4.5: capture sur l'affichage d'étiquetage

## Data set:

Index	id	title	link	tag	source	date_pub	date_save	intro	tag_code
0	352	Coronavirus : 186 nouve...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-13 16:32:00	2020-05-13 21:01:50.548...	ALGER-Cent- quatre-vingt...	5
1	353	Lutte contre le Covid-19: ...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-13 15:11:00	2020-05-13 21:01:50.560...	ALGER-Le ministre con...	5
2	354	Covid-19: le volume de pr...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 17:06:00	2020-05-13 21:01:50.565...	ALGER - Le volume de pr...	5
3	355	Covid-19: l'informatio...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 16:58:00	2020-05-13 21:01:50.574...	ORAN - Le ministre de ...	5
4	356	Covid-19: Djerad s'eng...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 16:44:00	2020-05-13 21:01:50.579...	RELIZANE - Le Premier mini...	5
5	357	L'Etat peut aider les di...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 16:08:00	2020-05-13 21:01:50.583...	ALGER - Le ministre de ...	5
6	358	Covid-19: 178 nouvea...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 16:24:00	2020-05-13 21:01:50.588...	ALGER - Cent soixante-sei...	5
7	359	Covid19: le GAAN propose...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 16:15:00	2020-05-13 21:01:50.593...	ALGER - Le Groupement a...	5
8	360	Benbouzid: il est indis...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 15:42:00	2020-05-13 21:01:50.604...	ORAN - Le ministre de ...	5
9	361	Covid-19: 7 millions de ...	http://www.aps.dz/s...	sante-	http://www.aps.dz	2020-05-12 14:48:00	2020-05-13 21:01:50.609...	ORAN - Le Premier mini...	5
10	362	Pluies orageuses su...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-13 17:30:00	2020-05-13 21:01:51.970...	ALGER-Des pluies, parf...	4
11	363	Ouargla/ Covid-19 : r...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-13 15:51:00	2020-05-13 21:01:51.975...	OUARGLA- Un premier lot ...	4
12	364	Alger : chute d'un a...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-13 11:07:00	2020-05-13 21:01:51.981...	ALGER- Les agents de la...	4
13	365	Prolongation du confine...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-13 10:50:00	2020-05-13 21:01:51.986...	ALGER- Les services de ...	4
14	366	L'infirmière Alcha ferhat...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-12 17:54:00	2020-05-13 21:01:51.992...	AIN DEFLA - Infirmière d...	4
15	367	Laghouat: Covid19: Cal...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-12 16:06:00	2020-05-13 21:01:51.997...	LAGHOUAT - Une opératio...	4
16	368	Ouargla: plus de 100 ...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-12 13:45:00	2020-05-13 21:01:52.002...	OUARGLA - Plus de 100 ...	4
17	369	Le Premier ministre ent...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-12 09:23:00	2020-05-13 21:01:52.006...	ORAN - Le Premier mini...	4
18	370	GECEAL: 45.000 tonne...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-11 17:19:00	2020-05-13 21:01:52.016...	ALGER- L'Etablissem...	4
19	371	Ghardaia : intensificat...	http://www.aps.dz/r...	regions	http://www.aps.dz	2020-05-11 13:04:00	2020-05-13 21:01:52.021...	GHARDAIA - La culture de l...	4
20	372	Lancement du concours du ...	http://www.aps.dz/c...	culture	http://www.aps.dz	2020-05-13 11:10:00	2020-05-13 21:01:53.874...	ALGER- Le Centre natio...	1

Figure 5.4.6 : capture du data set

## Split data into training and testing sets:

Il est important de conserver certaines données afin que nous puissions valider votre modèle. Pour cela, nous pouvons utiliser le `train_test_split` de Scikit Learn.

```
#2. Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(dataset['title'], dataset['tag_code'], random_state=1)

#-----
```

Figure 5.4.7 : capture de code spliting data

- **Convertir des résumés en vecteurs de nombre de mots :**
- Un classificateur Naive Bayes doit être capable de calculer combien de fois chaque mot apparaît dans chaque document et combien de fois il apparaît dans chaque catégorie. Pour rendre cela possible, les données doivent ressembler à ceci:
- [0, 1, 0,...]
- [1, 1, 1,...]

- [0, 2, 0,...]
- Chaque ligne représente un document et chaque colonne représente un mot. La première ligne peut être un document contenant un zéro pour «préchauffer», un pour «le» et un zéro pour «four». Cela signifie que le document contient une instance du mot «le», mais pas de «préchauffage» ou «four».
- Pour obtenir nos résumés dans ce format, nous pouvons utiliser CountVectorizer de Scikit Learn. CountVectorizer crée un vecteur de nombre de mots pour chaque résumé afin de former une matrice. Chaque index correspond à un mot et chaque mot apparaissant dans les résumés est représenté.
- Nous pouvons utiliser les arguments strip\_accents, token\_pattern, lowercase et stopwords pour exclure les non-mots, les nombres, les articles et d'autres choses qui ne sont pas utiles pour prédire les catégories de nos décomptes.

```
#.3 Convert abstracts into word count vectors
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(strip_accents='ascii', token_pattern=u'(?ui)\b\\w*[a-z]+\w*\\b', lowercase=True, stop_wor
X_train_cv = cv.fit_transform(X_train)
X_test_cv = cv.transform(X_test)
```

Figure 5.4.8: capture de code

### • view data :

Si nous souhaitons afficher les données et étudier le nombre de mots, vous pouvez créer un DataFrame du nombre de mots avec le code suivant:

```
#.3. view the data and investigate the word counts
word_freq_df = pd.DataFrame(X_train_cv.toarray(), columns=cv.get_feature_names())
top_words_df = pd.DataFrame(word_freq_df.sum()).sort_values(0, ascending=False)
```

Figure 5.4.9 : capture de code

affichage :



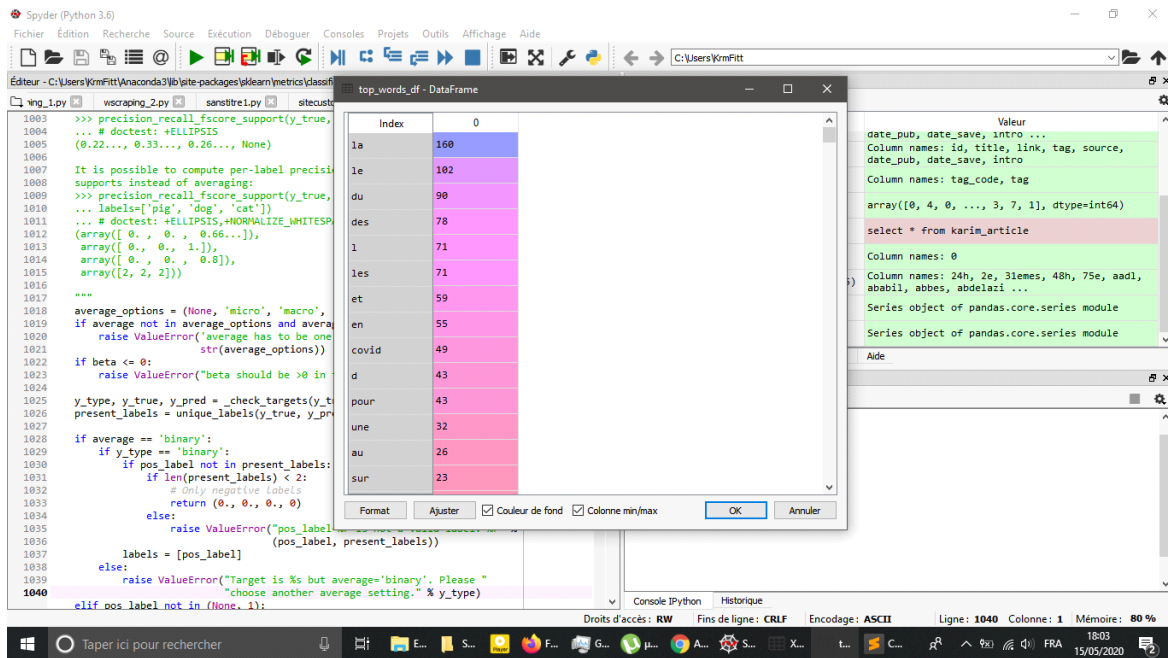


Figure 5.5.1: top words

- **Fitting the model :**

Nous sommes maintenant prêts à adapter un modèle de classifieur multinomial Naive Bayes à nos données d'entraînement et à l'utiliser pour prédire les étiquettes des données de test:

```
#4. Fit the model and make predictions
#Now we're ready to fit a Multinomial Naive Bayes classifier model to our training data and use it to predict the
from sklearn.naive_bayes import MultinomialNB
naive_bayes = MultinomialNB()
naive_bayes.fit(X_train_cv, y_train)
predictions = naive_bayes.predict(X_test_cv)
```

Figure 5.4.10 : capture de code

- **Checking results :**

```
#5. Check the results
#Let's see how the model performed on the test data:
from sklearn.metrics import accuracy_score, precision_score, recall_score
print('Accuracy score : ', accuracy_score(y_test, predictions))
print('Precision score : ', precision_score(y_test, predictions))
print('Recall score : ', recall_score(y_test, predictions))
```

Figure 5.4.11 : capture de code

- **Pour comprendre ces scores, il est utile de voir une ventilation:**

```
#To understand these scores, it helps to see a breakdown:
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
cm = confusion_matrix(y_test, predictions)
sns.heatmap(cm, square=True, annot=True, cmap='RdBu', cbar=False,
xticklabels=['sport', 'monde', 'culture', 'regions', 'economie', 'societe', 'sante-science-technologie'], yticklabel=
plt.xlabel('true label')
plt.ylabel('predicted label')
```

Figure 5.4.12 : capture de code

- Résultat :

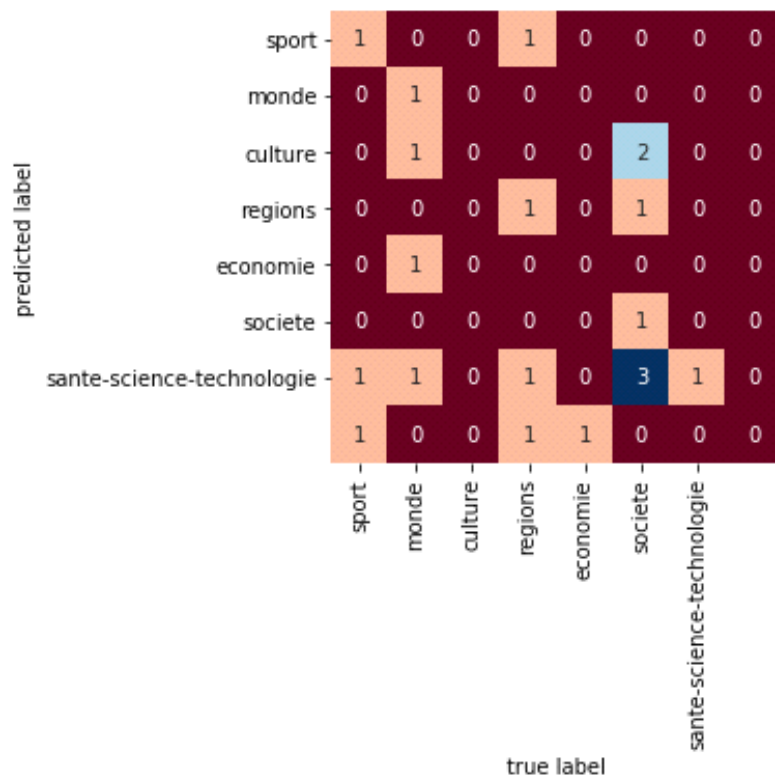


Figure 5.4.13 : capture sur le résultat

Le **score de précision** nous dit: sur toutes les identifications que nous avons faites, combien étaient correctes?

vrais positifs + vrais négatifs / total des observations:  $(18 + 19) / 40$

Le **score de précision** nous dit: parmi toutes les identifications éthiques que nous avons faites, combien étaient correctes?

vrais positifs / (vrais positifs + faux positifs):  $18 / (18 + 2)$

Le **score de rappel** nous dit: sur tous les vrais cas d'éthique, combien en avons-nous correctement identifiés?

vrais positifs / (vrais positifs + faux négatifs): 18 / (18 + 1)

## 6. Examiner les échecs du modèle

Pour rechercher les étiquettes incorrectes, nous pouvons placer les étiquettes réelles et les étiquettes prédites côte à côte dans un DataFrame.

Cet exemple est seulement pour deux catégories :

```
#6. Investigate the model's misses
#To investigate the incorrect labels, we can put the actual labels and the predicted labels side-by-side in a DataFrame.
sr_karim2=pd.Series( predictions);
sr_karim1=X_test.reset_index()
df_karim=pd.concat([sr_karim1,sr_karim2], axis=1)
df_karim.columns = ['index','title','tag_code']

df_karim['tag_name']=lb_make.inverse_transform(df_karim['tag_code'])

testing_predictions = []
for i in range(len(X_test)):

    if predictions[i] == 1:
        testing_predictions.append('culture')
    else:
        if predictions[i] == 0:
            testing_predictions.append('algerie')
check_df = pd.DataFrame({'actual_label': list(y_test), 'prediction': testing_predictions, 'abstract':list(X_test)})

check_df = pd.DataFrame.from_dict(testing_predictions, orient='index')
check_df.replace(to_replace=1, value='sport', inplace=True)
check_df.replace(to_replace=0, value='culture', inplace=True)]
```

Figure 5.4.14 : code for investigate incorrect labels

Résultat :

Index	abstract	actual_label	prediction
0	Consécration du principe ...	sport	sport
1	Guelma: 65% de taux de r...	4	sport
2	Formation professionne...	2	sport
3	Aéroport d'Alger: le ...	2	sport
4	Idir: une longue carri...	culture	culture
5	Nasri appelle à la...	2	sport
6	Participati...	culture	culture
7	ONU: le coronavirus ...	3	sport
8	Concours à distance du ...	culture	culture
9	Palestine: l'Unrwa récl...	3	sport
10	Tamazight: 20 titres no...	culture	sport
11	Covid-19: le confinement ...	sport	sport
12	La ministre de la Cultur...	culture	culture
13	Développeme...	2	sport
14	Hausse des prix/ volail...	2	sport
15	Confinement dans les hôt...	6	sport
16	Djerad: la mobilisation...	sport	sport
17	L'Etat peut aider les di...	5	sport
18	Le projet de révision de ...	sport	sport
19	Lutte contre le coronavir...	6	sport
20	Oran: 13 personnes gu...	4	sport

Figure 5.4.15 : capture sur l'affichage des prédictions

### 5.4.3 conclusion :